GBE

Genome Biology and Evolution

SMBE

Evidence for Deep Phylogenetic Conservation of Exonic Splice-Related Constraints: Splice-Related Skews at Exonic Ends in the Brown Alga *Ectocarpus* Are Common and Resemble Those Seen in Humans

XianMing Wu[1], Ana Tronholm[1,3], Eva Fernández Cáceres[1], Jaime M. Tovar-Corona[1], Lu Chen[2], Araxi O. Urrutia[1], and Laurence D. Hurst[1,*]

[1]Department of Biology and Biochemistry, University of Bath, Somerset, United Kingdom

[2]Human Genetics, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, United Kingdom

[3]Present address: Department of Biological Sciences, University of Alabama, Mary Harmon Bryant Hall, Tuscaloosa, AL

*Corresponding author: E-mail: l.d.hurst@bath.ac.uk.

Accepted: July 25, 2013

Abstract

The control of RNA splicing is often modulated by exonic motifs near splice sites. Chief among these are exonic splice enhancers (ESEs). Well-described ESEs in mammals are purine rich and cause predictable skews in codon and amino acid usage toward exonic ends. Looking across species, those with relatively abundant intronic sequence are those with the more profound end of exon skews, indicative of exonization of splice site recognition. To date, the only intron-rich species that have been analyzed are mammals, precluding any conclusions about the likely ancestral condition. Here, we examine the patterns of codon and amino acid usage in the vicinity of exon–intron junctions in the brown alga *Ectocarpus siliculosus*, a species with abundant large introns, known SR proteins, and classical splice sites. We find that amino acids and codons preferred/avoided at both 3′ and 5′ ends in *Ectocarpus*, of which there are many, tend, on average, to also be preferred/avoided at the same exon ends in humans. Moreover, the preferences observed at the 5′ ends of exons are largely the same as those at the 3′ ends, a symmetry trend only previously observed in animals. We predict putative hexameric ESEs in *Ectocarpus* and show that these are purine rich and that there are many more of these identified as functional ESEs in humans than expected by chance. These results are consistent with deep phylogenetic conservation of SR protein binding motifs. Assuming codons preferred near boundaries are "splice optimal" codons, in *Ectocarpus*, unlike *Drosophila,* splice optimal and translationally optimal codons are not mutually exclusive. The exclusivity of translationally optimal and splice optimal codon sets is thus not universal.

Key words: ESE, *Ectocarpus*, splicing, translational selection.

Introduction

Although for many years patterns of biased codon usage have been typically addressed in terms of translational optimality (and fit to the tRNA pool) (Duret 2002; Sharp et al. 2005), more recently the importance of exonic motifs involved in splicing has been seen to be relevant (Willie and Majewski 2004; Chamary and Hurst 2005; Parmley et al. 2006, 2007; Parmley and Hurst 2007; Warnecke et al. 2008). Chief among these motifs are exonic splicing enhancers (ESEs) (Blencowe 2000; Cartegni et al. 2002). At the RNA level, these motifs are responsible for the binding of SR proteins to the exonic parts of the unspliced RNA, thereby enhancing splicing at the neighboring exon–intron junction (Graveley 2000). In addition, they are responsible for retaining unspliced RNA in the nucleus (Taniguchi et al. 2007). Well-described ESEs in mammals—one of the few lineages where ESEs have been experimentally confirmed (Fairbrother et al. 2002, 2004; Fairbrother, Holste, et al. 2004; Ke et al. 2011)—are enriched toward the ends of exons (Fairbrother, Holste, et al. 2004), cause selective constraint at synonymous sites (Carlini and Genut 2006; Parmley et al. 2006), and have a highly skewed nucleotide usage,

typically being highly purine enriched (Tanaka et al. 1994; Fairbrother, Holste, et al. 2004; Parmley et al. 2007). Well-described ESEs occupy on average 30–40% of sequence near exon ends in mammals (Parmley et al. 2006). Note that as ESEs appear to be functional up to approximately 70 nt from an exon end (Fairbrother, Holste, et al. 2004), exons shorter than 140 bp can be considered to be all exon "end."

Owing to these three properties (high density, proximity to boundaries, and skewed nucleotide content), ESEs leave a marked footprint of codon usage near exon ends of mammalian genes, with codons more commensurate with involvement in ESEs (Parmley et al. 2007) being preferred near boundaries (Parmley and Hurst 2007). Similarly, when comparing synonymous codons, the one used more in ESEs is relatively preferred at exon ends over the synonym (Willie and Majewski 2004; Parmley and Hurst 2007). Thus, in mammals, although isochore composition is a strong driver of between-gene codon usage bias (Eyre-Walker and Hurst 2001), selection to preserve ESEs explains many of the intra–exon trends in codon bias. Amino acids also show skews in their usage as one approaches exon–intron junctions, with trends being well predicted by nucleotide content of ESEs and the codons that contribute to any given amino acid (Parmley et al. 2007). Indeed, comparing the usage of the 2-fold blocks of leucine and arginine with their respective 4-fold blocks supports the view that these trends are both the result of nucleotide-level effects and dominantly caused by splice-related constraints (Parmley et al. 2007). Just as knowing about ESEs makes sense of codon and amino acid trends, so too, conversely, *k*-mers that are enriched toward the ends of exons can be used to infer nucleotide preferences of splice-related motifs and to determine novel motifs (Lim et al. 2011) (N.B. codons are in frame 3-mers).

The trends seen in mammals have a series of further properties. For example, when usage trends at the 5′ and 3′ ends of exons are considered separately, it appears that the trends are largely symmetrical (Warnecke et al. 2008; Lim et al. 2011). That is, if a codon or amino acid is highly preferred at the 5′ end of exons, it is similarly highly preferred at the 3′ end. The logic of this symmetry is unclear, but it may accord with a model in which SR proteins aggregate on the ends of exons within the immature RNA and this aggregate defines, by the end of the cluster, a domain where the splice junction must reside. In such a model, there is no evident reason why different SR proteins should be under selection to bind 3′ and 5′ ends differently. However, such symmetry has to date only been observed in animals (Warnecke et al. 2008) and not in all of them. The 5′ ends of exons in *Caenorhabditis* worms, for example, are not simply different in composition to the 3′ ends; they show the opposite trends, that is, codons preferred at the 5′ ends are avoided at the 3′ ends and vice versa (antisymmetry). The 3′ end trends accord with the trends seen in all other taxa, with classical purine loading. The exceptional nature of worm's 5′ ends was hypothesized to reflect consequences of operonization in worm and the commensurate transplicing. The need to distinguish the 5′ ends of exons from the 5′ ends of genes, cut during transplicing, is suggested as the potential cause (Warnecke et al. 2008).

More generally, the trends in codon usage at the ends of exons in mammals correlate well with those seen in other animals, for example, *Drosophila* (Warnecke and Hurst 2007). This observation is important because *Drosophila*, unlike mammals, also has evident selection for use of "translationally optimal" codons, possibly to ensure mistranslation minimization (Akashi 1994; Drummond and Wilke 2008; Warnecke and Hurst 2010). In part, the cause of the strong correlation between end of exon usage in *Drosophila* and mammals reflects the fact that the "splicing optimal" set of codons and the "translationally optimal" set of codons are two almost mutually exclusive sets of codons, that is, translationally optimal codons tend to be those avoided near exon boundaries (Warnecke and Hurst 2007). At first sight, this mutual avoidance of the two sets seen in *Drosophila* makes some sense. If the two sets were the same, in highly expressed genes SR proteins would have difficulty binding exclusively to exonic ends, as all codons would be both translationally and splice optimal. Hence one might expect considerable splice disruption. Given such logic, it is worthwhile asking whether the same exclusivity rule applies in a very distantly related species.

Beyond *Drosophila*, whether the trends as observed in mammals are well conserved remains unclear, as the tendency to use SR proteins covaries with the intron density and size of introns (Warnecke et al. 2008). This trend possibly reflects an increased tendency toward exonization of splice site recognition as introns get ever larger, with small introns in a sea of large introns being the hardest to correctly splice using intronic information alone. At the other limit, a species such as *Saccharomyces cerevisiae* shows no preference trends (Warnecke et al. 2008), largely lacks SR proteins (Plass et al. 2008), and has very few and small introns. The nonanimal species previously analyzed (such as *Arabidopsis*) have very small introns and probably do not commonly use ESEs too, although SR proteins are possibly relatively ancient within eukaryotes but poorly described outside of the animal–fungal–plant crown group (Plass et al. 2008).

To examine whether the patterns seen in mammals might be relatively ancient requires analysis of distant genomes with abundant and relatively large introns. To this end, we selected for scrutiny the unusual genome of the brown alga *Ectocarpus siliculosus*. Brown algae share a common ancestor with the animal–fungal–plant crown group that predates the animal–fungal–plant common ancestor (Adl et al. 2005). The genome is well sequenced and annotated (Cock et al. 2010, 2012). It is unusual in being a nonvertebrate that is rich in introns (5.1 introns per kb of exon), and those introns tend to be large (mean intron size = 776 bp), meaning the genome is a strong candidate for one using ESEs and SR proteins to aid splicing,

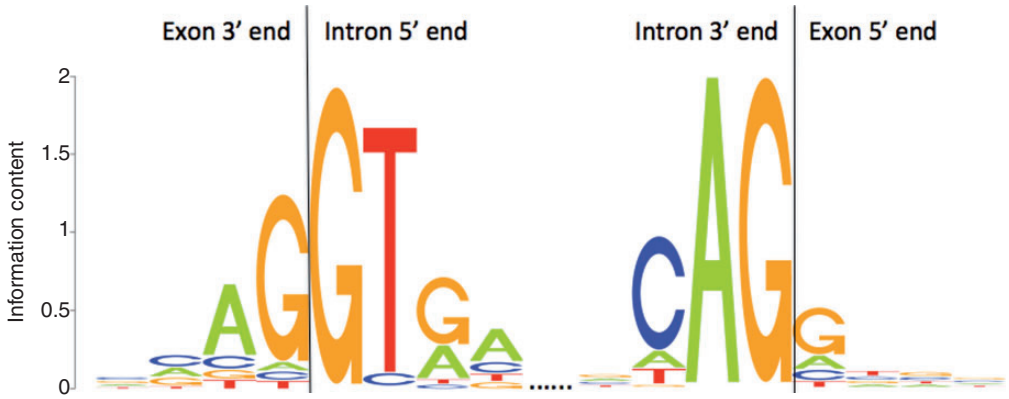

FIG. 1.—Splice site composition in *Ectocarpus*.

with a mean CDS size-to-gene size ratio of 0.27, comparable with mammals (Warnecke et al. 2008). As expected, annotation of the genome suggests it has SR proteins (Cock et al. 2010) (discussed later). The classical GT–AG rule applies in 95.3% of introns, the remainder being GC–AG introns (for sequenceLogo motifs, see fig. 1; for a longer span and evidence of a classical intronic 3′ polypyrimidine track, see supplementary fig. S1, Supplementary Material online). Importantly, much as with humans and other intron-rich genomes, but unlike some protists and intron-poor genomes (Irimia et al. 2007), there is not one hexameric motif that dominates intronic 5′ ends (GTGAGT at 12.5% is the most common). It thus appears an ideal candidate to ask whether the trends well resolved in humans are ancestral or animal specific. We also demonstrate that *Ectocarpus* has "translationally optimal" codons and thus ask whether these codons are never splice optimal codons.

Finally, taking advantage of what we discover to be some unusual features of the *Ectocarpus* genome, we reexamine the cryptic splice site avoidance model (Eskesen et al. 2004). This model posits that, with introns starting GT and exons ending in G, GGT should be avoided at the 3′ ends of exons (Eskesen et al. 2004) compared with the synonym GGC. *Ectocarpus* provides an unusually "clean" test of this prediction.

Materials and Methods

Establishing the Data Set for Analysis

The coding sequences (CDS) file and EMBL format exon information files for the brown alga *E. siliculosus* were downloaded from the database (http://bioinformatics.psb.ugent.be/genomes/view/*Ectocarpus*-siliculosus, last accessed September 16, 2013). The input CDS data were filtered to eliminate dubious sequences. We eliminated coding sequences that did not start with ATG, did not finish with a stop codon (TAA, TAG, and TGA), had internal stop codons, were not a multiple of three long, or contained one or more ambiguous nucleotides ("N"). In addition, those where the gene sequence length does not match the sum of the length of its exons as specified in the accompanying annotation files were eliminated. As we are interested in splice-related constraints, gene sequences that did not contain introns were also not examined. There are 16,579 coding sequences in the input file, of which 16,033 sequences qualified as suitable candidates.

Information of Expression Level of *Ectocarpus* Genes

The EST database of *Ectocarpus* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucest/?term=%22Ectocarpus±siliculosus%22[porgn%3A__txid2880], last accessed September 16, 2013). Using BLAST, we identified the number of ESTs associated with each gene (identity > 95%, *e* value < 0.01). The length-corrected EST hit rate (EST hits divided by the length of the gene) of each gene was regarded as the relative expression level of the gene.

HMMER Search for and Classification of SR Proteins

An SR protein reference data set, comprising 213 SR protein genes from different species, was established with the information from the website: http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node1.html (last accessed September 16, 2013). HMMER (Eddy 1998) was used to search for putative SR proteins in *Ectocarpus* genes (including those without introns) after multiple sequence alignment by MUSCLE.

To infer which, if any, of a set of cross species–conserved SR proteins our candidates might belong to, we performed a domain-based analysis, as previously described (Plass et al. 2008). In brief, we examined nine groups (families) of known SR proteins: SRp20 9G8, p54 SRp86, RY1, SC35-alias SRP1, SRm300, SRp30c-ASF, SRp40-55-75-alias SRP2, Topol-B, and Tra2. These were downloaded from http://www.bioinf.uni-leipzig.de/Leere/PRAKTIKUM/Protokolle/WS08/2/node6.html (last accessed September 16, 2013). We aligned, using MUSCLE, the different groups of proteins

separately. We then used "hmmbuild" of HMMER to make an "hmm" profile for each multiple sequence alignment. All profiles were collected to form a profile database. Using "hmmscan," we searched all candidate *Ectocarpus* proteins against the profile database. Finally, we determined the SR protein family that best matched each *Ectocarpus* SR candidate. To this end, we considered those domains within a given *Ectocarpus* protein that are in the same order as in the reference SR protein (these being the "collinear" domains). We then summed the score of collinear domain hits for any given *Ectocarpus* protein for each reference SR protein. To choose which family a given *Ectocarpus* protein belongs to, we selected the one whose sum score of collinear domain hits was highest. Finally, we accepted this classification if the sum score of the collinear hits for a multi domain protein, or a single hit for a single domain protein, was equal to or greater than 100.

Determining Trends in Amino Acid and Codon Usage

According to the information in the EMBL annotation files, we extracted every exon sequence for every qualifying gene. The trend in usage of each codon and amino acid was investigated as a function of the distance from the exon–intron boundary up to a distance of 34 codons (to accord with an earlier analysis [Warnecke et al. 2008]). Importantly, the codon in direct proximity to the boundary was eliminated, but was used to analyze splice site profiles. The 5′ and 3′ ends were considered separately. The first and last exons were excluded, leaving 95,331 exons. For each codon and amino acid under consideration, we determined the slope on the line of proportional usage across all exons, as a function of distance from the boundary and the Spearman rank correlation (ρ). A negative slope on the line, or a negative ρ, indicates a codon or amino acid that is preferred near exon ends, whereas a positive slope implies a codon or amino acid preferred at exonic cores and avoided at the ends. In previous analyses, codons preferred near exon ends were well predicted by the composition of experimentally defined ESEs (Parmley and Hurst 2007).

Human Exonic Splice Enhancer Data Sets

The majority of systematic attempts to define human ESEs use computational approaches, confirmed with experimental support. Typically, these approaches start with a presumption about that distribution of ESEs and look for the sequences most enriched in these trends. We analyze three such data sets. Fairbrother et al. (2002, 2004) presumed that ESEs will be enriched in exons compared with introns and more abundant in exons with weak splice sites than in those with strong splice sites. This is the RESCUE-ESE data set. Zhang and Chasin presumed ESEs will be enriched in internal noncoding exons of protein coding genes compared with unspliced pseudo-exons and 5′ untranslated regions. This is the PESE data set. Goren et al. (2006) looked for motifs that were more conserved than expected at synonymous sites and enriched compared with background codon usage rates. This is the ESR data set. In the latter case, a minority of the motifs were exonic splice inhibitors, the precise proportion being uncertain not least because ESEs can also function as exonic splice inhibitors depending on their position and context within the exon (Ke et al. 2011). The fourth data set we consider, Ke-ESE, derives from a purely experimental approach adopted by Ke et al. (2011). They considered the effects of all possible 4,096 6-mers at five locations in two model exons. Taking into account overlap sequences, this permitted the identification of numerous ESE hexamers.

We downloaded the ESR and Ke-ESE hexamers directly from the original papers. For the Ke-ESE set, we selected, as the authors did, the 400 hexamers with the highest scores. The RESCUE-ESE data set was downloaded from http://genes.mit.edu/burgelab/rescue-ese/ESE.txt (last accessed September 16, 2013), and the PESE original octamers were downloaded from http://www.columbia.edu/cu/biology/faculty/chasin/xz3/pese262.txt (last accessed September 16, 2013). PESE hexamers were extracted from octamers with a minimum of seven occurrences.

Assembling a Set of *Ectocarpus* Putative ESEs

The attempts to infer human ESEs have, as noted earlier, typically specified two criteria whereby ESEs are expected to be enriched (i.e., a ***R***elative ***E***nhancer and ***S***ilencer ***C***lassification by ***U***nanimous ***E***nrichment = RESCUE method). Here, we perform a similar RESCUE approach to define *Ectocarpus* ESEs. We consider that ESEs should be 1) enriched at exonic ends compared with introns and 2) that the usage of the ESE should increase from exon core to exon flank.

To determine the latter, for all 4,096 possible hexamers we considered their relative usage in exons, in all frames, as one moves away from exon ends. We considered only those exons longer than 160 bp to ensure that enrichment at exonic ends is truly such enrichment, rather than enrichment in short exons. The 5′ and 3′ ends were considered separately.

To consider those hexamers enriched at exon ends compared with intronic sequence, we considered exons longer than 100 bp and introns longer than 100 bp. We then considered the terminal 50 bp at each end of the exons and 50 bp at the end of the introns. For statistical analysis, it is important that there are the same number of introns as exons, so we randomly sampled from the larger data set to equalize the size of the two.

For each hexamer, we then considered its mean usage at exon ends and its mean usage at intron ends. We then calculated the difference in usage between the exon end and intron end. The 5′ exonic ends were compared with the 3′ intronic ends and vice versa. For each hexamer we can then define

$$\delta_{\text{Observed}} = \text{Exonic density} - \text{Intronic density}.$$

To consider the significance of this, we then pooled the relevant data from exons and introns, randomized them, and then considered the first half of the data as being pseudo-exon and the second half pseudo-intron. Repeating this 100 times for each hexamer we define

$$\delta_{\text{Pseudo}} = \text{Pseudo exonic density} - \text{Pseudo intronic density}.$$

A reasonable metric of the extent to which a given hexamer is enriched at exon ends compared with intronic ends is then:

$$Z = \frac{\delta_{\text{Observed}} - \overline{\delta_{\text{pseudo}}}}{\sigma_{\delta\text{Pseudo}}}$$

where $\overline{\delta_{\text{pseudo}}}$ is the mean of the hexamer usage in the 100 pseudo sets and $\sigma_{\delta_{\text{Pseudo}}}$ is the standard deviation in the usage across the pseudo sets. P was approximated by extrapolation from Z under an assumption of normality.

To generate a set of ESEs, we then considered those hexamers enriched in exon end compared with intron ($Z > 0$) and preferred near exon ends compared with core ($\rho < 0$, slope < 0), and then combined P values from the two approaches using Fisher's method. We then considered those hexamers with a combined $P < 0.05/4{,}096$ as putative ESEs.

CAI Calculation, Identification of "Optimal" Codons, and the Relationship with Gene Expression

A data set containing 43 ribosomal proteins was established and used as a reference "highly expressed" gene class. The codon usage in this set was analyzed using CodonW. We used this reference data set and the reference codon usage table from Codon Usage Database (www.kazusa.or.jp/codon/countcodon.html, last accessed September 16, 2013) to determine codon adaptation index (CAI) scores for all genes. To this end, we downloaded the local version of CAIcal, a CAI calculation program, from http://genomes.urv.es/CAIcal/ (last accessed September 16, 2013) and calculated the CAI values of all genes, except the ribosomal genes, which had been used for establishing the reference data set. The validity of the CAI index was examined by considering the relationship between expression level and CAI, again excluding the ribosomal genes. For any given synonymous block, the codon with the highest codon adaptation index was considered to be the "optimal" codon. tRNA copy numbers were obtained from http://plantrna.ibmp.cnrs.fr/ (last accessed September 16, 2013).

Alternative Splicing Event Calculation

Alternative splicing events in human and *Ectocarpus* genes were identified from 8,315,122 and 67,082 EST sequences, respectively, downloaded from the dbEST database (Boguski et al. 1993), using methods previously outlined (Chen et al. 2011). In brief, individual ESTs were matched to individual genes by aligning them to the genome sequence using GMAP (Wu and Watanabe 2005). Exon templates were then inferred from EST alignment coordinates. Alternative splicing events were identified by comparing alignment coordinates for each EST against the exon template.

Comparable alternative splicing event counts correcting for EST coverage were obtained using a transcript normalization protocol as described previously (Kim et al. 2007), where alternative splicing events per gene are calculated as the average number of alternative splicing events identified in 100 random samples of 10 ESTs.

Results

Ectocarpus Has Multiple SR Proteins

Before asking whether binding of SR proteins to ESEs leaves a footprint of biased codon and amino acid usage in proximity to intron–exon boundaries, we first established the profile of SR proteins within the genome. To search for candidates, we did a HMMER search, training the HMMER on an established collection of SR proteins. In total, we identified 54 putative SR proteins, including three previously annotated as SR proteins (Esi0638_0002, Esi_0327_0029, and Esi0164_0021) (supplementary result S1, Supplementary Material online). In the original build of the genome, a further putative SR protein was identified (Esi0089_0034; annotated as "splicing factor, arginine/serine-rich 2, RNAp interacting protein, putative"). This was not identified by HMMER. Although many of the extra hits are unlikely to be SR protein (e.g., eukaryotic translation initiation factor 3′, subunit a), several more have suggestive RNA binding functions. Most of the extra hits are not annotated.

To clarify just which SR proteins the HMMER search might have revealed, we performed an additional domain-based analysis, comparing our proteins with nine SR proteins that are relatively well conserved through plants, animals, and fungi (Plass et al. 2008). We found robust evidence for 18 of the *Ectocarpus* genes being members of 1 of 5 SR protein families (fig. 2; supplementary table S1, Supplementary Material online). This includes the three previously annotated SR proteins (supplementary table S1, Supplementary Material online). We conclude that *Ectocarpus* has a good number of SR proteins but probably not the full set described in humans (Long and Caceres 2009).

A High Proportion of Amino Acids and Codons Show Preference/Avoidance Trends

To determine which, and how many, codons and amino acids show significantly skewed usage in proximity to exon–intron junctions, we considered the relative usage of all codons and amino acids as a function of distance from an exon–intron junction, ignoring the codon in immediate proximity to the junction. We examined the 3′ and 5′ ends separately. Any codon or amino acid preferred near a boundary will have a negative slope and a negative Spearman rank correlation (ρ)

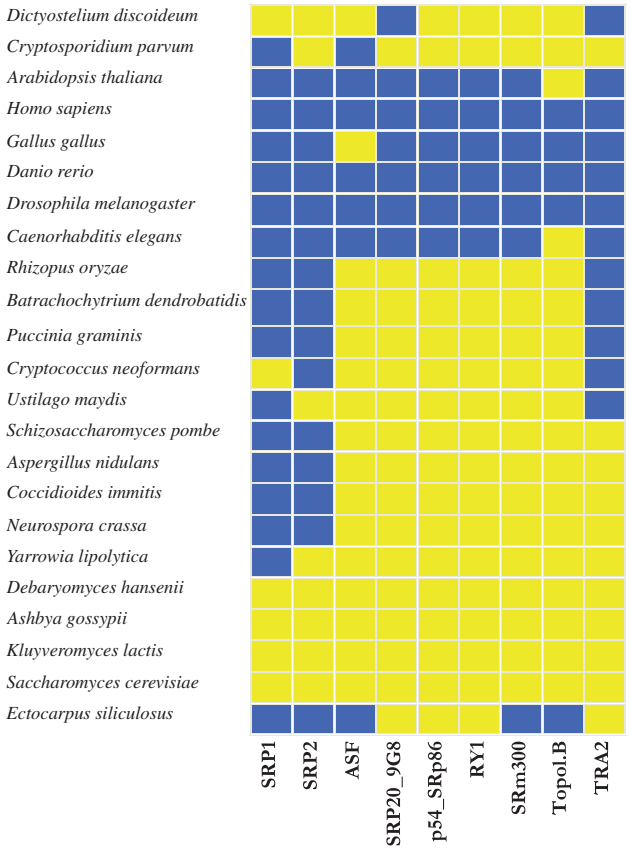

Fig. 2.—Presence or absence of SR and SR-related proteins in *Ectocarpus* compared with other taxa. Presence of at least one member of a gene family in a given species is indicated in dark blue or absence in light yellow. Data for all species bar *Ectocarpus* from Plass et al. (2008).

between its relative usage and the distance from the boundary. A positive slope/ρ score indicated avoidance near a boundary relative to usage more core to exons. The slope/ρ values, we considered to be measures of the preference/avoidance trends.

Most codons and amino acids showed significant preference/avoidance directions near exon–intron boundaries (supplementary tables S2 and S3 and figs. S2 and S3, Supplementary Material online). Before Bonferonni correction, 86% of codons and 96% of amino acids showed significant trends ($P < 0.05$) at the 5′ exonic ends, and 88% of codons and 91% of amino acids showed significant trends ($P < 0.05$) at the 3′ exonic ends. After correction, these numbers dropped to 68% of codons and 83% of amino acids showing significant trends at the 5′ exonic ends, and 69% of codons and 83% of amino acids showing significant trends at the 3′ exonic ends. Overall, considering all codons with at least one synonym, 66% of comparisons showed significant trends after Bonferonni multitest correction, and 83% of amino acids analyses showed significant trends.

These figures compare strikingly with what has been seen before. A priori we expect that species with relatively small

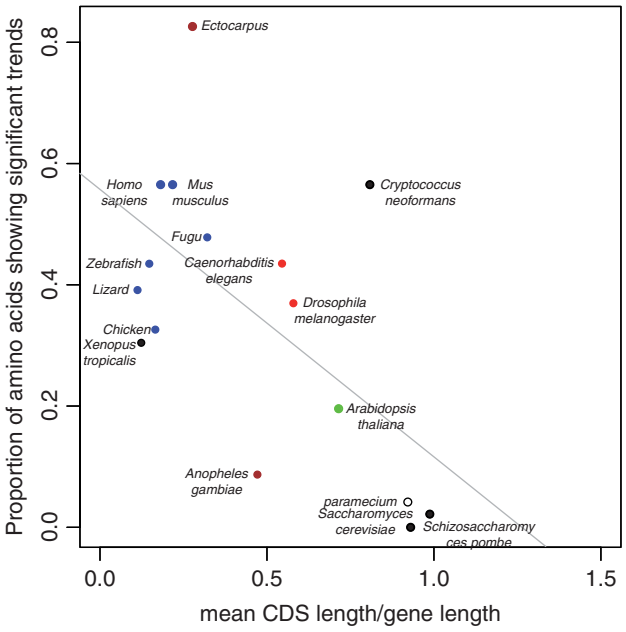

Fig. 3.—The proportion of amino acids showing significant preference/avoidance trends after Bonferroni correction as a function of the average ratio of mature CDS to gene length across multiple species.

exons sitting in an intron-rich sea will be those that will be under selection for adding exonic splice information to bolster the intronic signals. This should in turn be reflected in more codons and more amino acids showing skewed usage near boundaries. This supposition is generally supported by the finding that species in which the ratio of the mature CDS size to gene size is small are those in which a greater proportion of amino acids or codons show significant skews toward exon ends (fig. 3). When we consider our new data in this light, while *Ectocarpus* certainly has a low CDS-to-gene ratio, the proportion of amino acids showing a skew remains unusually high (fig. 3). We note that this cannot be an artifact of sample sizes (the higher the sample size, the more likely significant skews will be seen even if trends are weak), as humans and mice have fewer significant trends but more exons analyzed.

Preference/Avoidance Trends at 3′ and 5′ Exon Ends Are Similar

Is the pattern of symmetry seen in most animals, but not so far reported outside of animals, also seen in *Ectocarpus*? To address this, we considered for each amino acid and each codon the trend in its usage approaching the 5′ and 3′ ends of exons. The slope on this line and the Spearman ρ values were considered. We then considered the correlation between the figures when comparing the 5′ and 3′ ends. We found that overall exons tend to be symmetric, with a strong correlation

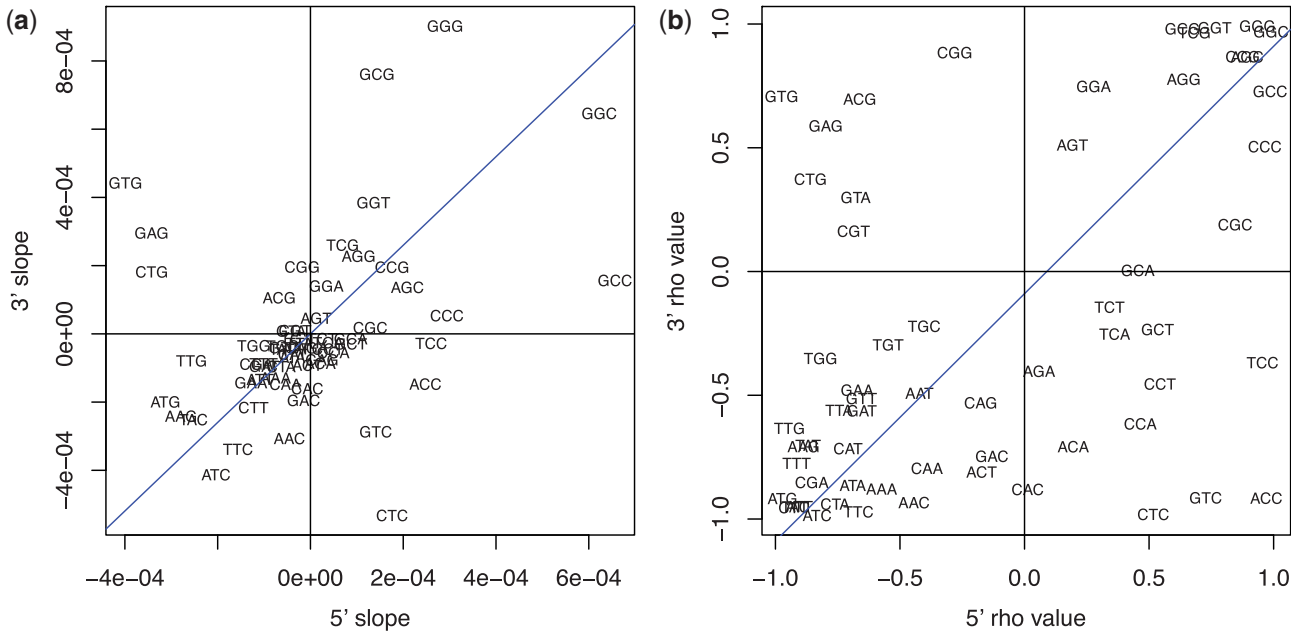

Fig. 4.—Examination of symmetry of preference/avoidance trends for codons. For both 5′ and 3′ exon ends, we considered (*a*) the slope on the line of relative usage versus distance from the boundary ($\rho = 0.40$, $P = 0.0014$) and (*b*) the Spearman rank correlation for the same comparison ($\rho = 0.47$, $P = 0.00014$). A negative slope or a negative ρ indicates a codon that is preferred near an exon boundary. For each codon, we can then compare these trends at the 5′ and 3′ ends. We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

both between the slopes and the ρ values for codons (fig. 4: from binomial test with success as preservation of the sign of the slope, $P = 0.004$; from Spearman rank correlation on slopes, $P = 0.001$) and amino acids (fig. 5: from binomial test with success as preservation of the sign of the slope, $P = 0.0026$, from Spearman rank correlation on slopes, $P = 0.002$).

However, closer scrutiny indicates a further nuance, namely, that C and G ending codon usage can be antisymmetrical (fig. 4). CTC, GTC, and ACC are all highly disfavored at exonic 5′ ends but highly preferred at 3′ ends, whereas GTG, GAG, ACG, CGG, and CTG all show the opposite pattern. Generally, at 4-fold degenerate sites, C and G show somewhat antisymmetrical patterns, with C avoided at 5′ ends while G, although highly abundant, avoided at 3′ ends, meaning as one approaches the boundary, its usage declines (fig. 6*a* and *b*). However, these patterns are not simple enough to be explained solely in terms of C or G content, as many C ending codons, including CCC, are symmetrical. Nonetheless, we conclude that the symmetry rules are not universally respected.

Usage Trends Near Exon Junctions in *Ectocarpus* Resemble Those Seen in Humans

To consider whether the trends seen in humans are also seen in a very distant species with large intronic content such as *Ectocarpus*, we considered how the ρ values for each amino acid, 5′ and 3′ resembled those seen in humans. We can then ask two questions. First, do the trends overall correlate between the two species? Second, what proportion of comparisons has a conserved direction of preference/avoidance between the two species?

To this end, we considered the correlations between the trends (ρ values) seen in the two species, considering 5′ and 3′ ends separately. Remarkably, despite over 1 billion years of divergence, we found that at both 5′ and 3′ exonic ends the trends correlate well (5′ end: $\rho = 0.68$, $P = 0.0005$; 3′ end: $\rho = 0.53$, $P = 0.01$; fig. 7 and table 1). At the 3′ ends, only 4 of the 23 codon blocks (we treated 2-fold and 4-fold blocks differently here) did not have a conserved preference trend (binomial test, $P = 0.002$). At the 5′ ends, although the trends are correlated, seven codon blocks had different trends, rendering the result nonsignificant (binomial test, $P = 0.09$). However, when we restricted analysis to those amino acids showing significant trends (before Bonferroni correction), we observed that at both the 5′ and 3′ ends, the proportion conserved was significant (table 1). Moreover, we noted that of the four amino acids that were antisymmetrical in *Ectocarpus* (E, R, V, and T), three (E, V, and T) were also antisymmetrical in humans (supplementary fig. S4, Supplementary Material online).

We can perform a similar analysis using codons (excluding ATG and TGG owing to lack of synonyms). Again we found strong concordance between trends seen in humans and those seen in *Ectocarpus* (5′ end: $\rho = 0.50$, $P = 5.17 \times 10^{-5}$; 3′ end $\rho = 0.58$, $P = 1.7 \times 10^{-6}$; fig. 8 and table 1). The conservation patterns mirrored what we saw at the amino acid

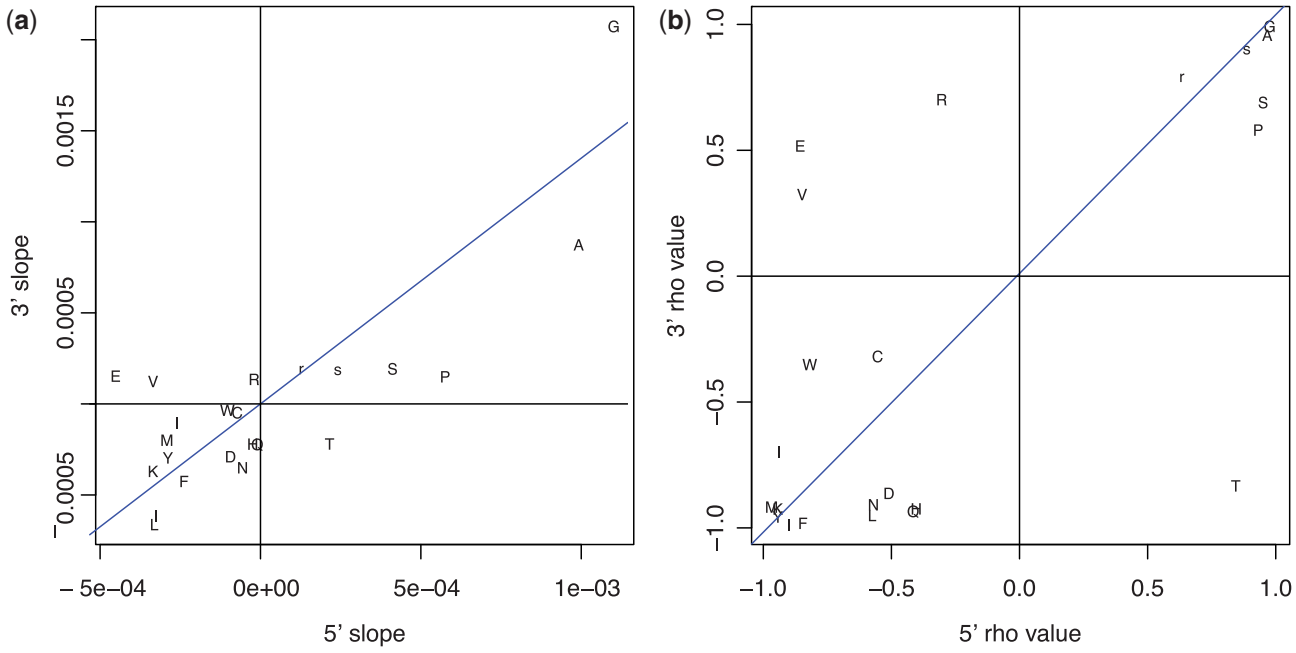

Fig. 5.—Examination of symmetry of preference/avoidance trends for amino acids. For both 5′ and 3′, we considered both the slope on the line of relative usage versus distance from the boundary and the Spearman rank correlation for the same comparison. For each amino acid, we can then compare these trends at the 3′ and 5′ ends, considering either (*a*) slope ($\rho = 0.60$, $P = 0.003$) or (*b*) ρ ($\rho = 0.68$, $P = 0.0005$). We note that overall exons tend to have symmetrical trends. The blue line indicates the SMA regression.

level. At the 3′ ends, we saw a strong correlation, and only 12 showed reverse trends (binomial test, $P = 5.13 \times 10^{-6}$). At the 5′ ends, the effects were more modest. A significant correlation was observed, but of 59 codons, 23 showed reverse trends at the 5′ ends ($P = 0.12$). As above, restricting analysis to only those codons showing significant trends, at both 5′ and 3′ ends, more than expected show conservation of direction (table 1). Nucleotide usage at 4-fold degenerate sites was also comparable between *Ectocarpus* (fig. 5*a* and *b*) and humans (fig. 5*c* and *d*), although in *Ectocarpus* the C and T preferences at the 3′ end were more similar than seen in humans. Overall, these results suggest a deep phylogenetic conservation of splice-associated trends in amino acid composition as one approaches exonic ends, most especially at the 3′ ends.

Ectocarpus Has Low Rates of Alternative Splicing

The similarity that we see between humans and *Ectocarpus* in terms of which codons and amino acids are preferred and avoided near boundaries suggests that the selection on the nucleotide usage in DNA or RNAs at exon ends is for similar reasons. Why then does *Ectocarpus* have so many more amino acids and codons showing significant trends (fig. 2)? The metric we use is by no means perfect, as it is sensitive to sample sizes (number of exons examined). However, high numbers of trends seen for *Ectocarpus* compared with mouse/human cannot be an artifact of sample sizes, as the sample sizes in vertebrates (in terms of number of exon ends) are larger than those in *Ectocarpus*.

One possibility that explains the large number of skews in *Ectocarpus* is that alternative splicing might be relatively rare in *Ectocarpus*. The consequence of this would be that most exons are consistently under strong selection to be spliced correctly. By contrast, if in humans many exons are splicing errors (Zhang et al. 2009), then we would not expect strong selection to preserve ESEs in all exons. The uniformity of splice sites in *Ectocarpus* (fig. 1) would be consistent with the hypothesis that most exons are under selection to be properly spliced.

Preliminary data suggest that alternative splicing is indeed rare in *Ectocarpus*. A detailed examination of splice forms has been performed on one gene family, the cytosolic glutathione transferases. While 11 genes were identified, only one had an alternative transcript (Franco et al. 2008). Although this is much lower than the rate seen in humans, in whom nearly all intron-bearing genes have at least two isoforms (Pan et al. 2008), this difference might reflect, at least in part, differences in the depth of study (Brett et al. 2002).

To compare alternative splicing levels in both humans and *Ectocarpus* allowing for depth of EST sequencing, we performed two tests. First, we measured alternative splicing levels in genes from both species after transcript number normalization. For this, alternative splicing per gene was measured as the average number of alternative splicing events detected in 1,000 random samples of 10 ESTs. We obtained

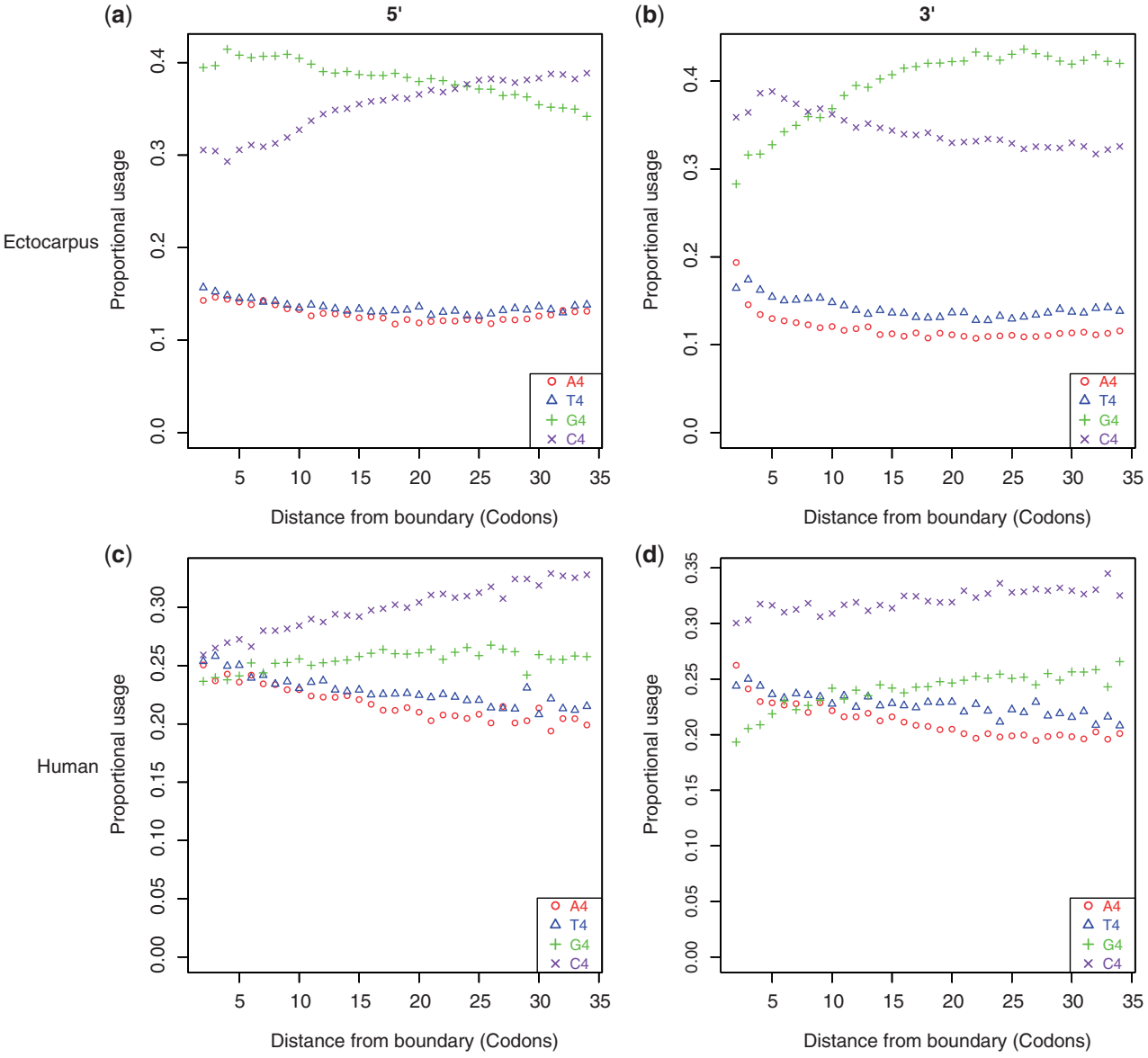

Fig. 6.—Nucleotide usage at 5′ and 3′ exon ends at 4-fold degenerate sites in *Ectocarpus* and humans. The data here use only exons longer than 64 codons so that all exons contribute equally at all distances. The plots are (*a*) *Ectocarpus* 5′ end, (*b*) *Ectocarpus* 3′ end, (*c*) human 5′ end, and (*d*) human 3′ end.

this comparable index of alternative splicing for 8,772 human genes and 69 *Ectocarpus* genes. We found that while *Ectocarpus* genes had an average of 0.41 events per gene (median of 0), human genes had an average of 5.35 events per gene (median of 4.55). This difference is highly significant (*t*-test, $P = 2.15 \times 10^{-68}$). Second, we compared the average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. When genes were divided according to their average number of aligned ESTs, the average number of alternative splicing events per gene was considerably higher for humans compared with *Ectocarpus* at all nine EST per gene counts ($P = 0.004$ from binomial test: $N = 4{,}861$ human; 326 *Ectocarpus*, fig. 9). We conclude that in *Ectocarpus* alternative splicing is rare compared with that seen in humans. Although this is consistent with the possibility that alternative transcription rates might impact on the net skew in nucleotide usage, this hypothesis requires considerable further cross-taxon analysis.

Ectocarpus Putative Exonic Splice Enhancers Resemble Those Seen in Humans

Above we compared human and *Ectocarpus* exonic ends as regards trends in codon usage. The trends seen in

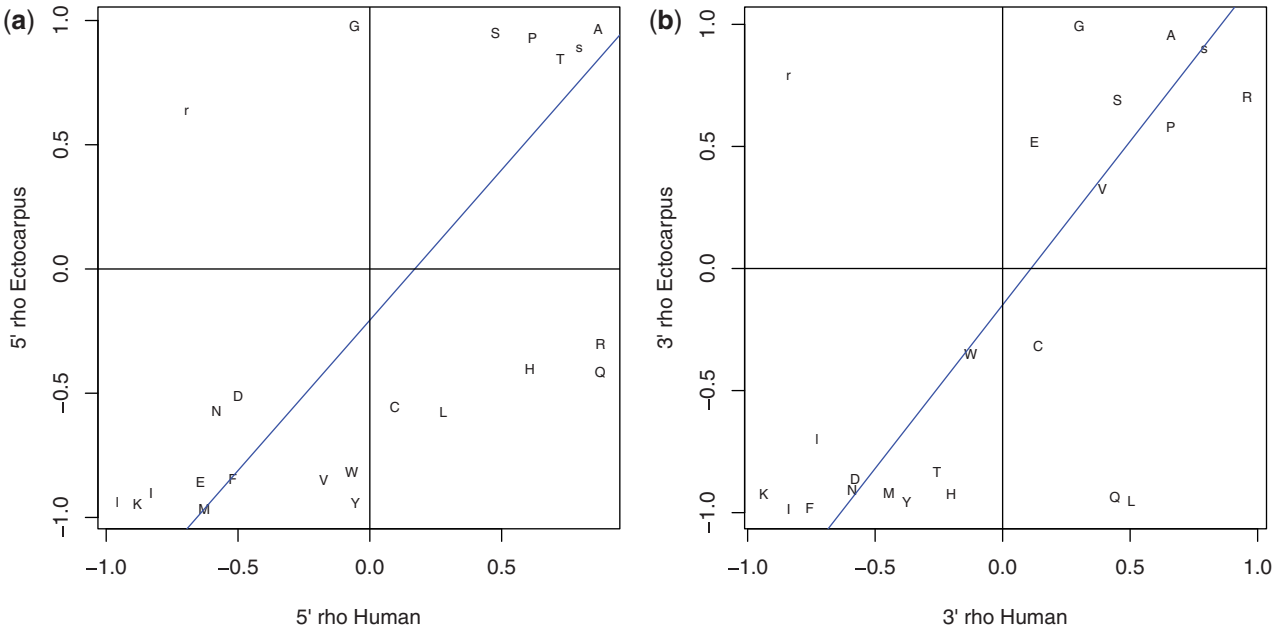

Fig. 7.—Comparison of preference/avoidance trends at the amino acid level between humans and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by ρ (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at (*a*) 5′ ($\rho = 0.68$, $P = 0.0005$) and (*b*) 3′ ($\rho = 0.53$, $P = 0.01$) ends of exons are shown. The blue line is the SMA regression line.

Table 1
Conservation of Trends between Humans and *Ectocarpus*

	Binomial Test		Spearman's Rank Correlation	
	Changed Direction	***P***	**ρ**	***P***
All observations				
5′ AA	7 from 23	0.093	0.6779	0.0005
3′ AA	4 from 23	0.0026	0.5316	0.0100
5′ codon	23 from 59	0.1175	0.5017	5.17E−05
3′ codon	12 from 59	5.13E−06	0.5773	1.70E−06
Significant observations				
5′ AA	3 from 16	0.021	0.7559	0.0011
3′ AA	3 from 17	0.013	0.5000	0.0430
5′ codon	11 from 43	0.0019	0.5694	6.75E−05
3′ codon	5 from 38	4.26E−06	0.5938	8.51E−05

Note.—The trends in usage of all codons and amino acids at 3′ and 5′ ends of exons were compared between humans and *Ectocarpus*. Two reporting statistics were considered. First, we ask using a binomial test whether the proportion of observations changing direction of trend is different from expected under a null where trends are free to evolve. Second, we consider a Spearman rank correlation test. As the analysis can be biased by considering trends that are very marginal and nonsignificant, we perform a second analysis where only significant trends ($P < 0.05$ before Bonferroni correction in both species) are used.

mammals reflect the nucleotide content of ESEs (Parmley and Hurst 2007). This is to be largely expected, as ESEs are hexamers that tend to be enriched at exon ends in any frame and codons are 3-mers in frame and hence are likely to be nonindependent of ESE-imposed trends. ESEs are, however, not described in *Ectocarpus* so we cannot perform the same analysis. We can, however, attempt to determine which hexamers might function as ESEs and compare this set of candidates with those identified in humans.

To this end, we asked of *Ectocarpus* 1) which hexamers are enriched at exonic ends compared with intronic sequence and 2) which hexamers are used at exon ends more than in exon centers (i.e., have a negative slope of proportional usage against distance from exonic end). We then considered as candidate ESEs the hexamers most enriched on both axes. Note that this method is far from perfect in so much as we also identified causes of skew in codon usage probably not related to ESEs but rather to avoidance of cryptic splice sites (discussed later).

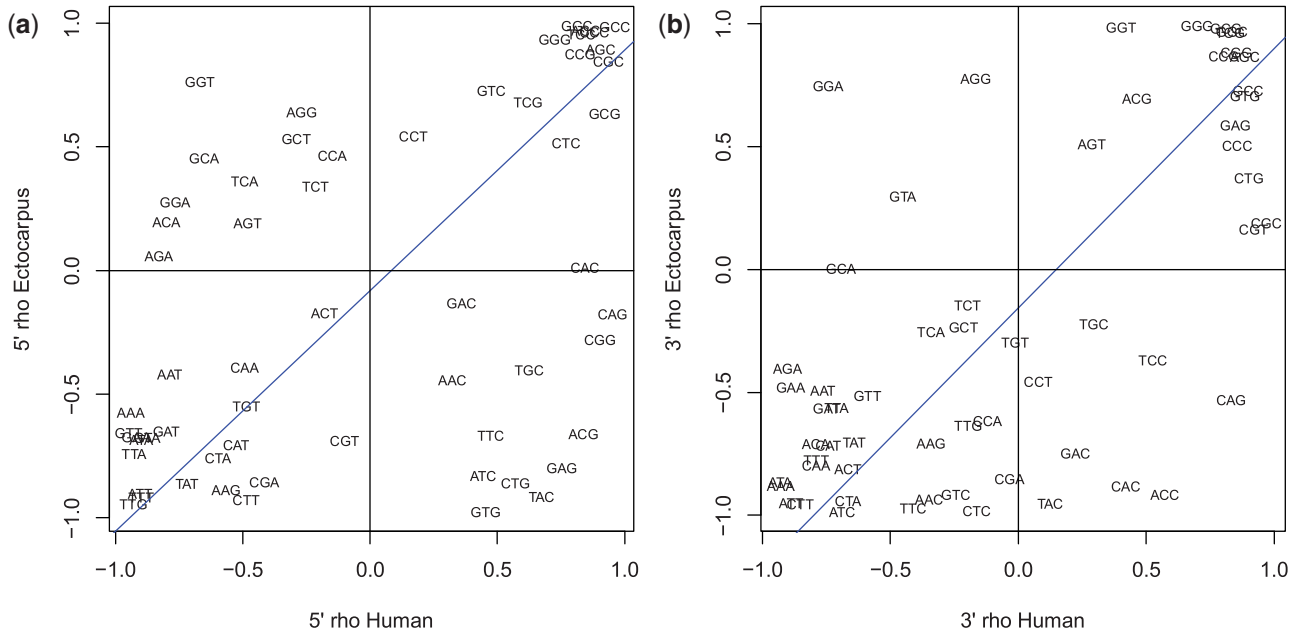

Fig. 8.—Comparison of preference/avoidance trends at the codon level between humans and *Ectocarpus*. The amino acid level preference avoidance trends, assayed by ρ (the rank correlation of proportional usage of the amino acid to distance from an exon boundary), at (*a*) 5′ ($\rho = 0.50$, $P = 5.17 \times 10^{-5}$) and (*b*) 3′ ($\rho = 0.58$, $P = 1.7 \times 10^{-6}$) ends of exons are shown. The blue line is the SMA regression line.

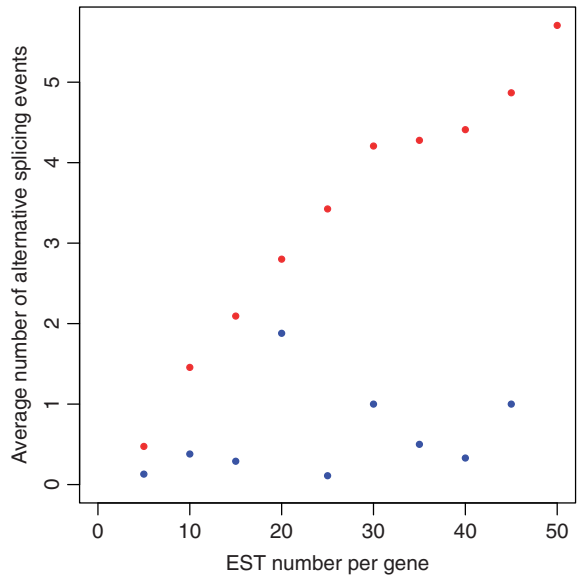

Fig. 9.—The average number of alternative splicing events detected when genes are grouped according to the number of ESTs aligning to them. Data for humans in red; data for *Ectocarpus* in blue.

We identified 904 3′ ESEs and 919 5′ ESEs (supplementary table S4, Supplementary Material online). The 5′ and 3′ hexamers are different from each other. We observed 189 in common but by chance, we would expect 203. Moreover, while the 5′ hexamers are, like classically described SR protein-binding ESEs, highly purine enriched (A + G = 64.4%), the 3′ set are, if anything, pyrimidine enriched (A + G = 42%), which is consistent with the C enrichment at 4-fold sites at the 3′ exonic ends (fig. 6). The set of 189 ESEs that are common to 5′ and 3′ ESEs are purine rich (A + G = 59.3%).

There are four high-throughput data sets attempting to identify ESEs in humans (see Materials and Methods). Unfortunately, these four have remarkably few hexamers in common—just 10 of more than 900 putative hexamers are found in all four data sets. We consider those hexamers found in at least 3 of the 4 data sets as being a robust set of human ESEs ($N = 54$). For both our 3′ and 5′ set of hexamers, we found considerably more overlap than expected under a null model in which the human set of ESEs and the *Ectocarpus* set were assumed to be independent. For the 5′ ESEs, we expected 12 hexamers in common but observed 39, more than 8 standard deviations than expected by chance ($P \ll 0.0001$). For the 3′ end ESEs, the effect was more modest but still highly significant: we observed 26 in common between the two sets, where less than 12 are expected by chance, nearly 5 standard deviations than expected by chance ($P < 0.0001$). Of the 189 hexamers that are in common at 5′ and 3′ ends, 18 are also in the set of 54 human ESEs while fewer than 3 are expected under a random null. This deviation is almost 10 standard deviations from expectations ($P \ll 0.0001$). All of these degrees of concordance between *Ectocarpus* and humans are considerably

greater in magnitude than the concordance witnessed between some of the initial four human data sets. We conclude that despite the unusual base composition of 3′ ESEs in *Ectocarpus*, there is a significant resemblance between human ESEs and *Ectocarpus* ESEs. The trends, especially those seen at the 5′ ends, are consistent with a deep and strong phylogenetic conservation of SR protein-binding preferences.

Translationally Optimal and Splice Optimal Codons Are Not Mutually Exclusive in *Ectocarpus*

In *Drosophila*, the set of codons enriched near exon ends accords with those commensurate with ESEs, and correlate well with the trends seen in mammals (Warnecke and Hurst 2007). These splice optimal codons are very different from the translationally optimal codons, with just one codon being in both sets (Warnecke and Hurst 2007). Is this mutual exclusivity also seen in *Ectocarpus*? To address this, we first must ask whether *Ectocarpus* is like *Drosophila* in having a translationally optimal class of codons. To this end, we first examined codon usage in the ribosomal proteins, these being the most highly expressed genes. Given the difference in the codon usage in the ribosomal genes and the codon usage in the genome as a whole, we could then ascribe each gene a CAI score. We then ask whether, excluding the ribosomal protein training set, the more highly expressed genes show higher CAI. There is a weak but significant correlation between CAI and expression level (Pearson correlation, $r = 0.084$; $P = 2.6 \times 10^{-13}$, supplementary fig. S5, Supplementary Material online).

In addition, we compared the optimal codons, as defined by over usage in ribosomal proteins, for each synonymous set with the tRNA copy numbers (assuming these to be a rough guide to tRNA levels) and asked if the optimal codon within each block was also the one with the most abundant tRNA. In 12 of 18 synonymous blocks, this was the case (supplementary table S5, Supplementary Material online). By randomizations, involving extracting at random two codons from each synonymous block, we asked how often we expected by chance to see 12 of 18 matching, given the structure of the genetic code. In 100,000 simulations we observed 12 or more matches in less than 1,000 incidences. We conclude that the optimal codons tend to be those matching the more abundant tRNAs ($P < 0.001$). *Ectocarpus* is, in this regard, more like flies than mammals, and is under translational selection.

Given the above result, we can now address whether the translationally optimal codons might be different from the splice optimal codons. To define splice optimal codons, we consider all those preferred near exon boundaries (at both 5′ and 3′ ends) that are significantly skewed after Bonferonni correction ($P < 0.05/118$ at both ends and $\rho < 0$) (supplementary table S5, Supplementary Material online). This defines 16 codons, although some of these are from the same codon block. Indeed, of 18 amino acids with more than one synonym, 10 amino acids have no splice preferred codons. In the remaining cases, three amino acids have all their codons as splice optimal (F, I, K, and Y). In three (H, L, and R) of the remaining four informative cases the translationally optimal codon, defined by reference to usage in ribosomal proteins, is not a splice optimal codon, but in K it is. To examine the significance of this, we considered a simulation in which we define for each of the four codon blocks the number of splice optimal codons and randomly sampled that number out of the number of codons in the block. We then ask how often the pseudo-splice set of codons and the pseudo translationally optimal codon matches. We then considered how often we see 1 or fewer matches. We found that we expect this to happen about 41.6% of the time, thus there is no evidence that splice optimal and translational optimal codons are under selection to differ.

We can be less stringent and define a splice optimal codon as any codon showing preference toward any exon end (not both 5′ and 3′ ends) after Bonferonni correction ($P < 0.05/118$). This gives 34 splice optimal codons (supplementary table S5, Supplementary Material online). There are only four potentially informative synonymous codon blocks in which some but not all of the codons are splice optimal. As regards translational optimality defined in terms of usage in ribosomal proteins, N and T have splice optimal codons that are translational optimal ones, whereas Q and R have the opposite. Again we see no significant evidence that splice optimal and translational optimality are divergent (from simulation: $P = 0.65$).

Additionally, for each codon block we can ask which codon is the most splice preferred. This we define as the codon with the most significantly negative slope using both 5′ and 3′ analyses. If no codon has a significantly negative slope then we consider the one with the most negative slope to be the splice preferred codon. We find that in 9 of 18 incidences the splice preferred codon is also the translationally optimal codon. We reject the hypothesis that splice optimal codons tend not to be translationally optimal codons (by simulation: $P = 0.91$).

Evidence of Cryptic Splice Site Avoidance

The nucleotide composition at 3′ exonic ends, allows us to provide an unusually "clean" test of the cryptic splice site avoidance model (Eskesen et al. 2004). Given that introns start GT and end AG, to avoid cryptic splice sites, it is argued (Eskesen et al. 2004) that AG residues should be avoided near 5′ ends of exons and GT should be avoided at the 3′ ends. One difficulty with any such analysis in mammals, however, is that, nucleotide usage in ESEs trends to go in the same direction as predictions from the cryptic splice avoidance model (Chamary and Hurst 2005). *Ectocarpus* provides an opportunity to test the cryptic splice model as at the 3′ ends both T and C are weakly preferred and show very similar relative trends (fig. 6*b*). As exons tend to end G in the majority of

incidences (fig. 1), the cryptic splice avoidance model thus predicts that at 3′ exon ends GGT should be avoided compared with GGC (a cryptic splice could occur between the two G residues in GGT), but [A|C|T]GT need not be avoided compared with [A|C|T]GC. Precise expectations for [A|C|T]GC and [A|C|T]GT are not however clear, their relative usage potentially reflecting background nucleotide trends. Given this, we asked solely whether GGT/GGC behaves differently from [A|C|T]GC and [A|C|T]GT, with the latter three consistent in their behavior. We observed just this, with profound avoidance of GGT compared with GGC, but preference for [A|C|T]GT, compared with [A|C|T]GC, near boundaries at the 3′ ends of exons (fig. 10). Note that both GGN and CGN are 4-fold degenerate codons so this comparison is especially well controlled. At the 5′ ends of exons the pattern is reversed with GGT being preferred over GGC, which is to be expected given the overall nucleotide composition, C being strongly avoided 5′ and T being weakly preferred. In sum, the preference of GGC over GGT at exonic 3′ end is consistent with the cryptic splice site model.

At the exonic 5′ end as introns end AG and exons commonly start with a G (fig. 1), the cryptic splice model predicts that AGG should be avoided compared with AGA. This is observed (supplementary fig. S6, Supplementary Material online). This test is not a strong one however as, while exons regularly end G, the preference to start with a G is weaker.

Although the cryptic splice model makes good sense of the preference for GGC over GGT at 3′ ends, most other trends cannot be explained in terms of splice avoidance. For example, preference for [A|C]GT over [A|C]GC most probably reflects processes acting more generally. We presume, as typically done (Lim et al. 2011), that most of the trends observed reflect the nucleotide content of splice motifs, such as ESEs.

Discussion

The analysis of the *Ectocarpus* genome has provided the first insight into the splice-related forces operating in a very distant relative of vertebrates in a species with intronic content comparable with that of vertebrates. The extent to which the trends observed in vertebrates accord with those seen in *Ectocarpus* are striking given the vast evolutionary distance between the two groups. It seems, therefore, parsimonious to presume that this reflects splice-related constraints, most probably the conservation of the binding motifs of SR proteins, not least because trends seen in humans accord well with those expected given the nucleotide composition of ESEs (Parmley and Hurst 2007; Parmley et al. 2007). Moreover, that *k*-mer enrichment in the vicinity of exon boundaries is a successful method to identify new splicing motifs, supports the supposition that the codon trends that we observe reflect splice-related motifs (Lim et al. 2011). The correspondence between our predicted set of *Ectocarpus* ESE hexamers and human ESEs is notable. It is to be expected that, just as vertebrate ESEs can function in fungi (Webb 2005), so they might function in brown algae too. Nonetheless, given our results regarding the cryptic splice site avoidance model, we see no reason to suppose that ESE enrichment is the sole cause of all the trends that we observe.

Why *Ectocarpus* has so many codons and amino acids showing strong preference avoidance trends (and also so many putative ESEs) is unclear. The possibility that alternative splicing is rare in *Ectocarpus*, hence resulting in selection on most exons most of the time for correct splicing, is consistent the data but requires further scrutiny. As regards the trends seen at the amino acid level, an alternative possibility to splice related selection is that we are detecting preference for one amino acid above another, owing to the hypothesized tendency of protein modules to reside in individual exons, as conjectured by the introns-early hypothesis (Gilbert et al. 1986). Aside from the fact that the one-module one-exon hypothesis is probably untenable (Stoltzfus et al. 1994; Logsdon 1998), this possibility is rejected in humans, not least because of the 6-fold degenerate amino acids, two (L and R) show opposite trends within the 2-fold and 4-fold degenerate blocks, these trends being well predicted by involvement in ESEs (Parmley et al. 2007). Similarly, for the 2-fold block of arginine (R) in *Ectoparpus*, at the 5′ exon ends both codons are avoided, whereas the three of the four codons of the 4-fold block are preferred. Within the 4-fold blocks of both valine and threonine there are both codons that are significantly avoided and significantly preferred. Similarly, within the 4-fold degenerate block of arginine and the 2-fold degenerate glutamic acid at 3′ exon ends one of the two is significantly avoided and one is significantly preferred. These differing trends within synonymous codon blocks support the hypothesis that, at least in part, the trends that we observed are owing to nucleotide level, not protein level, effects. Nonetheless, most pairs of codons in 2-fold degenerate blocks have preference trends in the same direction. This could reflect either some relationship to protein structure or similar splice-related selection (e.g., ESE involvement) owing to the first two bases in the 2-fold degenerate codons being identical in the synonyms.

Unlike *Drosophila*, *Ectocarpus*, while having evidence of being under translational selection, shows no evidence of selection to make splice optimal and translationally optimal codons distinct. Why might the two genomes differ? One possibility is that selection for translational optimality is that stronger in *Drosophila*. Indeed in *Ectocarpus* the correlation between CAI and expression level is rather weak. Were this weakness real (as opposed to an artifact of limited and noisy expression data) then selection to force divergence between translationally optimal and splice optimal codons may also be weak. Another possibility is that the observation

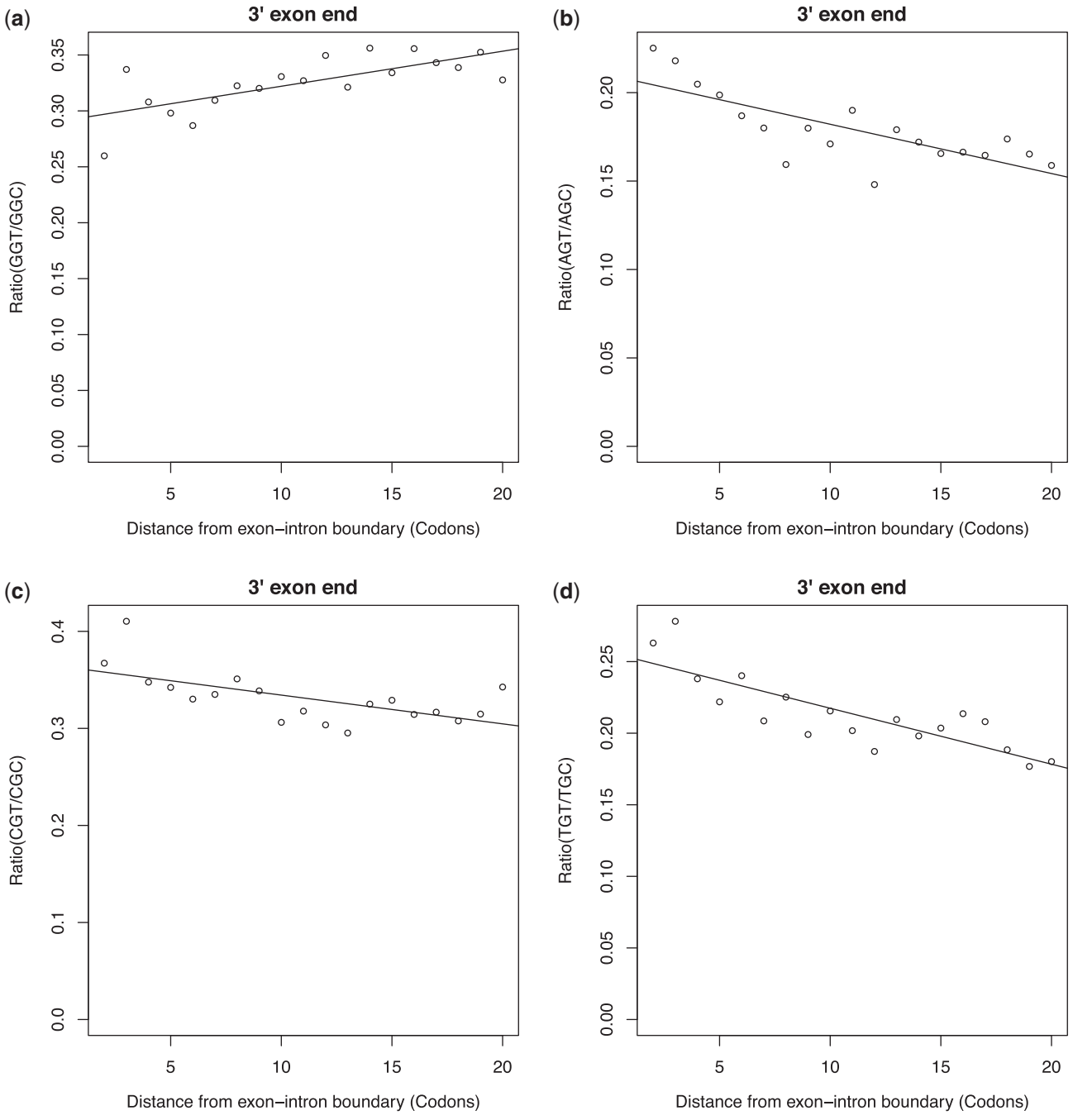

Fig. 10.—Relative usage of NGT against NGC at synonymous sites at exonic 3′ ends (*a*) N = G, (*b*) N = A, (*c*) N = C, (*d*) N = T.

in *Drosophila*, while seemingly having an attractive explanation, is an accidental consequence of selection on translational optimality and splice optimality happening to go in opposite directions. Until it is better understood why certain codons end up being translationally optimal this issue will be hard to resolve. Nonetheless, we can now provide an exemplar where translational optimality has not obviously selected on a set of codons distinct from the splice optimal set.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Simon Dittami for advice on expression resources. A.O.U. is a Royal Society Dorothy Hodgkin

Research Fellow and L.D.H. is a Royal Society Wolfsom Research Merit Award Holder. This work was supported by the a CONACyT scholarship (to J.M.T.-C.), the University of Bath (to X.W.), and the Erasmus program (to E.F.C.).

Literature Cited

Adl SM, et al. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Eukaryot Microbiol. 52: 399–451.

Akashi H. 1994. Synonymous codon usage in *Drosophila* melanogaster: natural selection and translational accuracy. Genetics 136:927–935.

Blencowe BJ. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci. 25: 106–110.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST—database for "expressed sequence tags". Nat Genet. 4:332–333.

Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. 2002. Alternative splicing and genome complexity. Nat Genet. 30:29–30.

Carlini DB, Genut JE. 2006. Synonymous SNPs provide evidence for selective constraint on human exonic splicing enhancers. J Mol Evol. 62: 89–98.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 3:285–298.

Chamary JV, Hurst LD. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet. 21:256–259.

Chen L, Tovar-Corona JM, Urrutia AO. 2011. Increased levels of noisy splicing in cancers, but not for oncogene-derived transcripts. Hum Mol Genet. 20:4422–4429.

Cock JM, et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. Nature 465:617–621.

Cock JM, et al. 2012. The *Ectocarpus* genome and brown algal genomics the Ectocarpus Genome Consortium. In: Piganeau G, editor. Genomic insights into the biology of algae. Amsterdam (The Netherlands): Elsevier. p. 141–184.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341–352.

Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Eskesen ST, Eskesen FN, Ruvinsky A. 2004. Natural selection affects frequencies of AG and GT dinucleotides at the 5′ and 3′ ends of exons. Genetics 167:543–550.

Eyre-Walker A, Hurst LD. 2001. The evolution of isochores. Nat Rev Genet. 2:549–555.

Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol. 2:E268.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. 2002. Predictive identification of exonic splicing enhancers in human genes. Science 297: 1007–1013.

Fairbrother WG, et al. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Res. 32:W187–W190.

Franco P-Od, Rousvoal S, Tonon T, Boyen C. 2008. Whole genome survey of the glutathione transferase family in the brown algal model *Ectocarpus siliculosus*. Mar Genomics. 1:135–148.

Gilbert W, Marchionni M, McKnight G. 1986. On the antiquity of introns. Cell 46:151–154.

Goren A, et al. 2006. Comparative analysis identifies exonic splicing regulatory sequences—the complex definition of enhancers and silencers. Mol Cell. 22:769–781.

Graveley BR. 2000. Sorting out the complexity of SR protein functions. RNA 6:1197–1211.

Irimia M, Penny D, Roy SW. 2007. Coevolution of genomic intron number and splice sites. Trends Genet. 23:321–325.

Ke S, et al. 2011. Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. 21:1360–1374.

Kim E, Magen A, Ast G. 2007. Different levels of alternative splicing among eukaryotes. Nucleic Acids Res. 35:125–131.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A. 108:11093–11098.

Logsdon JM. 1998. The recent origins of spliceosomal introns revisited. Curr Opin Genet Dev. 8:637–648.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. Biochem J. 417:15–27.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 40:1413–1415.

Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol. 23:301–309.

Parmley JL, Hurst LD. 2007. Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Mol Biol Evol. 24:1600–1603.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5: 343–353.

Plass M, Agirre E, Reyes D, Camara F, Eyras E. 2008. Co-evolution of the branch site and SR proteins in eukaryotes. Trends Genet. 24:590–594.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33:1141–1153.

Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF. 1994. Testing the exon theory of genes—the evidence from protein-structure. Science 265:202–207.

Tanaka K, Watakabe A, Shimura Y. 1994. Polypurine sequences within a downstream exon function as a splicing enhancer. Mol Cell Biol. 14: 1347–1354.

Taniguchi I, Masuyama K, Ohno M. 2007. Role of purine-rich exonic splicing enhancers in nuclear retention of pre-mRNAs. Proc Natl Acad Sci U S A. 104:13684–13689.

Warnecke T, Hurst LD. 2007. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. Mol Biol Evol. 24:2755–2762.

Warnecke T, Hurst LD. 2010. GroEL dependency affects codon usage-support for a critical role of misfolding in gene evolution. Mol Syst Biol. 6:340.

Warnecke T, Parmley JL, Hurst LD. 2008. Finding exonic islands in a sea of non-coding sequence: splicing related constraints on protein composition and evolution are common in intron-rich genomes. Genome Biol. 9:r29.

Webb CJ. 2005. Exonic splicing enhancers in fission yeast: functional conservation demonstrates an early evolutionary origin. Genes Dev. 19:242–254.

Willie E, Majewski J. 2004. Evidence for codon bias selection at the pre-mRNA level in eukaryotes. Trends Genet. 20:534–538.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875.

Zhang Z, et al. 2009. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. BMC Biol. 7:23.

Associate editor: John Archibald